

The local information content of the protein structural database

Srikar Rao, Qing-Lin Zhu, Sandor Vajda and Temple Smith

BioMolecular Engineering Research Center and Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received 16 December 1992; revised version received 22 March 1993

A simple study of the information available within proteins of known structure reveals both the limits of structure prediction based on overall statistical correlations between sequence and structure, and the areas where there is still a possibility for further structural database exploitation.

Secondary structure; Information content; Probability; Homolog

1. INTRODUCTION

Under suitable conditions, most proteins adopt a unique three-dimensional conformation determined by the amino acid sequence. In spite of its limited size, the database of known protein structures provides a wealth of information on the relationships between sequences and corresponding folded structures. Given the extreme complexity of any *de novo* prediction approach to complete tertiary structures, numerous heuristic methods have been devised to exploit these data in estimating limited structural features such as the locations of secondary structure elements. Most secondary structure prediction begins with the statistical assumption that within the known structures, the occurrence of the different amino acids is significantly correlated with their occurrence in different secondary structure elements.

It is now generally recognized that over the past 18 years little progress has been made in improving the accuracy of statistical methods beyond the 60% or so obtained early on. Even very recent attempts to use adaptive statistical approaches, such as neural nets, have shown only limited improvement [1–3]. Some success was achieved when one restricted or conditioned such analyses to the protein being a member of a particular structural class [4]. However, the overall limit on prediction accuracy appears to remain when one includes the uncertainty in the prediction of a protein's structural class.

Statistical methods of secondary structure prediction exploit only local information and necessarily neglect non-local interactions between residues far removed along the sequence but close to each other in space. The goal of this study is to delineate the inherent limitations on secondary structure prediction using such local information. The local information content is evaluated

in 117 non-homologous proteins of known structure and their close homologs. We find that there exists an upper limit of 64% on the information, given only the correlation between a single amino acid and the secondary structure at the same position and a neighboring position. Such correlations between amino acid type and secondary structure are the basis for the Chou–Fasman method in the form of propensities, whereas correlations between secondary structure assignments of neighboring residues were indirectly incorporated through the employment of ‘window averaging’ and thresholds. Our analysis further shows that the apparent limit on prediction accuracy can be significantly raised if: (i) nearest neighbor information is used for both the amino acids and the structural states, (ii) homologous sequence information is used; or (iii) such information is combined with explicit information on the length and connectivity of secondary structural elements in real proteins. The first case has recently been demonstrated by Vajda [5] for short peptides; the second case indirectly by the many successful applications of structural modeling by homology [6] (Levin, J.M., Pascarella, S., Argos, P., Garnier, J., *Protein Eng.* (in press) and more directly by Zvelebil et al. [7]; and the last case through the work of Stultz et al. [8].

2. EXPERIMENTAL

A set of 117 non-homologous proteins was selected from the Brookhaven PDB release 47 [9] after generating all pairwise maximally similar sequence alignments among all PDB sequences (the obvious, close homologues, such as those having the same name, function, species or size were excluded at the outset) with the dynamic programming algorithm of Smith and Waterman [10]. These 117 sequences were then aligned with five or more of their close homologues in the SWISSPROT database (release 21) using the method of Smith and Smith [11]. To estimate the additional information given by these homologs, we recorded only the distinct amino acid types occurring at each aligned position. These in turn were used to estimate association frequencies of amino acid types and secondary structures.

Each position in these multiple alignments was assigned a structural state by two different methods: first by that of Kabsch and Sander [12];

Correspondence address: T.F. Smith, BMERC, Boston University, 36 Cummings St., Boston, MA 02215, USA.

and second by direct assignment of one of eight regions in the plane of backbone dihedral angles (see Fig. 1). The rectilinear orientations of these regions were chosen to align with the apparent linear relationship between Φ and Ψ , seen most clearly in the traditional α -helical region and its mirror image region for $\Phi > 0$ (see Fig. 1 legend). With the exception of the 't' (turn) and the 'c' (coil) designation used by Kabsch and Sander [12], there is a high correlation between the two structure state assignments (data not shown). The 17,794 aligned positions with structural state assignment and the associated list of distinct amino acids at each position were used to calculate a wide range of conditional probabilities.

The primary results of this study can be presented in terms of a Shannon information measure [13]. The average missing information is:

$$\langle MI \rangle = -\sum_k P(e_k) \log_2 [P(e_k)]$$

where $P(e_k)$ is the probability of event e_k , and MI is expressed in 'bits'. The sum is over all possible events for which the probability distribution is defined. $\langle MI \rangle$ is a measure of the average information obtained upon observation of the actual events, given prior knowledge of the probability distribution, $P(e)$, only. For conditional probabilities this measure takes on the form:

$$\langle MI \rangle = -\sum_j P(e_j) \sum_k P(e_k | e_j) \log_2 [P(e_k | e_j)]$$

where $P(e_k | e_j)$ is the probability of event e_k given that the event e_j has occurred. For example, given no information other than that there are eight possible outcomes (with a priori equal probability), the observation of the actual structure at each residue position would reduce the 'missing' information by three bits. On the other hand, knowledge of the conditional probability of observing a particular structural state given the amino acid at that position (Chou-Fasman-like propensities) has less information missing prior to the observation of the actual structure; hence the observation of the actual structure leaves less information to be gained.

We present this missing information as a relative measure, taking as our reference the prior knowledge of the number of structural states only.

3. RESULTS

Table I presents the information measures for a range of statistical knowledge available in our aligned PDB data set. As already mentioned, if all the structural states are equally likely, i.e. $P(S_k) = 1/8$, then there are 3 bits of missing information. Since the conformational states have unequal probabilities, the missing information is less, $\langle MI \rangle = 2.46$ bits. Perhaps most surprisingly, the statistical knowledge of this probability of each structural state being observed independent of which amino acid occupies that position (line 2, Table I), is nearly as informative as the knowledge of the probability $P(S_k | a_k)$ of the association of each amino acid with each such structural state (line 4, Table I). Thus contrary to expectation, the correlation between amino acid type and structural state provides very little (0.08 bits) additional information. A major gain in information is obtained from the combination of prior knowledge of both the statistical associations of the 20 amino acids with each structural state and the nearest neighbor association between structural state types, as reflected in the conditional probability $P(S_k | a_k \& S_{k-1})$ of observing state S_k at position k given the amino acid a_k at that position and the preceding structural state S_{k-1} at the neighboring position $k-1$.

If the probabilities $P(S_k | a_k \& S_{k-1})$ are known, the missing information is reduced to 1.09 bits in the Kabsch and Sander case. Thus we have about 64% of the originally missing 3 bits of information required to uniquely specify one out of the eight Kabsch and Sander

Table I
Measures of local information content of the protein structural database

Prior knowledge probabilities	Average missing information, $\langle MI \rangle$		Gain in $\langle MI \rangle$ relative to the reference information ^a	
	Kabsch	Φ - Ψ	Kabsch	Φ - Ψ
$P(S_k) = 1/8$ for all structural states S	3.00	3.00		
$P(S_k)^b$	2.46	2.30	0.54	0.70
$P(S_k S_{k-1})$	1.15	1.89	1.85	1.11
$P(S_k a_k)^b$	2.38	2.11	0.62	0.89
$P(S_k a_k \& a_{k-1})$	2.19	1.97	0.81	1.03
$P(S_k a_k \& S_{k-1})$	1.09	1.67	1.91 (64%)	1.33 (44%) ^c
$P(S_k a_k \& S_{k-2})$	1.52	1.84	1.48	1.16
$P(S_k a_k \& S_{k-3})$	1.89	1.91	1.11	1.09
$P(S_k a_k \& S_{k-4})$	2.06	1.96	0.94	1.04
....				
$P(S_k a_k \& S_{k-8})$	2.27	2.04	0.73	0.96
....				
$P(S_k a_k \& S_{k-12})$	2.31	2.06	0.69	0.94
$P(S_k a_k \& a_{k-1} \& S_{k-1})$	0.83	1.27	2.17 (72%)	1.73 (58%)
$P(S_k a_k \& a_{k'}^d)$	1.44	1.14	1.56 (52%)	1.86 (62%)
$P(S_k S_{k-1} \& a_k \& a_{k'}^d)$	0.40	0.79	2.60 (86%)	2.21 (74%)

^a Percentages are shown in parentheses.

^b We have used a shorthand notation, where the subscript, k , refers only to the relative sequence positions.

^c Values of 62% are obtained using a reduced state space composed of only the seven highly occupied regions shown in Fig. 1.

^d The symbol $a_{k'}$ indicates any second distinct homolog residue at the position k , in addition to the residue a_k of the protein of known structure.

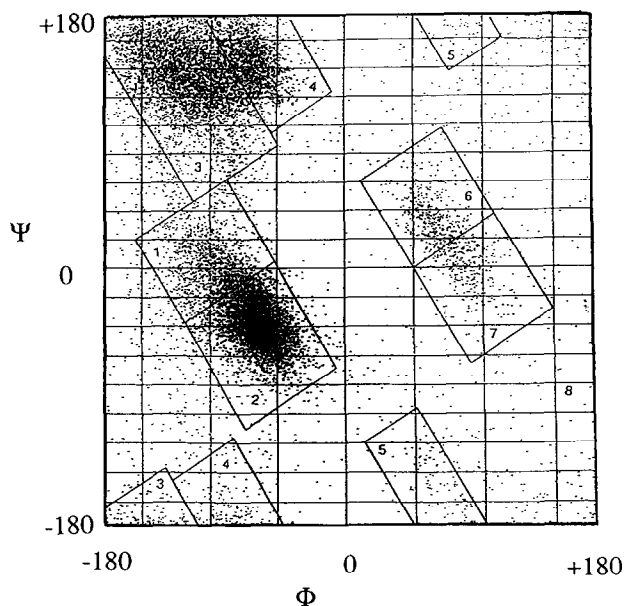


Fig. 1. A plot of the Φ - Ψ angles for the 117 sequences in the Brookhaven Protein Data Bank used in this study. The rectangles indicate the regions used to define our eight discrete states. These are marked as 1 and 2, the traditional α -helical regions; 3 and 4, the traditional β -sheet regions; and 6 and 7, the traditional turn or loop regions. Region 5 has no simple interpretation in terms of secondary structure. All of the remaining region is defined as state 8 to allow for proper normalization of the probabilities.

secondary structural states at each residue position! Note that in the eight Φ - Ψ states case, only 44% of the information appears to be recovered by knowledge of both the amino acid and the neighboring state on average. It would appear that in defining various secondary structures, additional information has been included. This is of course correct, particularly in the case of turns where at least two of the discrete Φ - Ψ states are combined in a particular order for any tight turn class. However, this is somewhat misleading since, for completeness, the eighth state was defined as all Φ - Ψ points not included in any of the seven high density regions in Fig. 1. If we use only the seven highly populated regions and re-normalize the probabilities, nearly 62% of the information is resolved.

4. DISCUSSION

Depending on the definition of structural states, 60–65% may be considered as an upper limit on what is obtainable by any structural prediction scheme having as its sole input the knowledge of the amino acid structural preferences and some knowledge of the probability of the continuation of any structural element, expressed here as the probability $P(S_k | a_k \& S_{k-1})$. This includes simple window averaging algorithms where, by requiring an average over three or four residues to exceed a given value, one is in effect not allowing second-

ary structures of shorter total length. This value derived here is an upper limit since the conditional probabilities used assume that the preceding neighboring state is known, while in any prediction scheme it is at best estimated.

Since the correlation between neighboring states can be expected to extend over a greater range than just nearest neighbors, we have also calculated a number of longer range structural state conditional probabilities. The information gain can be seen to drop off rapidly for both the KS states and the discrete Φ - Ψ states with the distance. The auto-correlations have an average 'half' distance of just under three residues. This is a little surprising given that hydrogen bonding in helices and tight turns extends over such a distance at a very minimum and few β -strands are less than three residues.

While the data set of known non-homologous protein structures is still limited statistically, it is sufficient to calculate a number of higher order correlations. For example, we can calculate the conditional probability $P(S_k | a_k \& a_{k-1} \& S_{k-1})$ of observing the state S_k (k indicates the position), given the amino acid type a_k at that position, and both the neighboring structure S_{k-1} and the neighboring amino acid type a_{k-1} . This appears to be able on average to provide up to 72% of the missing three bits. Lim [14] first proposed using such data in 1974, but before the required structural information was available. Even today this data is rather sparse in that many of the 24,320 ($20 \times 19 \times 8 \times 8$) conditional probabilities are estimated at zero from the observed frequencies. Given that there are, in principle, no such absolute zeroes, (approximated by assuming single occurrences for all zeroes in the frequencies) all estimated probabilities would be slightly closer to uniformity, thereby reducing the estimates on missing information. Vajda [5] has recently exploited such sparse data in the prediction of probable peptide structure distributions, with some success. However, part of the success was obtained by not restricting oneself to a single state prediction.

Fig. 2 contains the distributions of cardinalities where cardinality for an amino acid is defined as the number of distinct amino acids observed at the same position in the aligned sequences. From these distributions it is clear that the probability of being in a conserved position is very different for the different amino acids. This suggests that one should be able to exploit such information and its correlations with structure. Note that the amino acids, Ala, Ile, Leu, Val and Asp, have distributions for this positional degeneracy that differ little from that given by a simple Poisson model. The amino acid aligned position cardinalities in plot C of Fig. 2 are quite different. However, more important differences may be associated with the structural associations. For example, while Ile is very rarely found in an aligned position containing no other amino acids (as a conserved residue), when it is conserved, it is five times more likely to

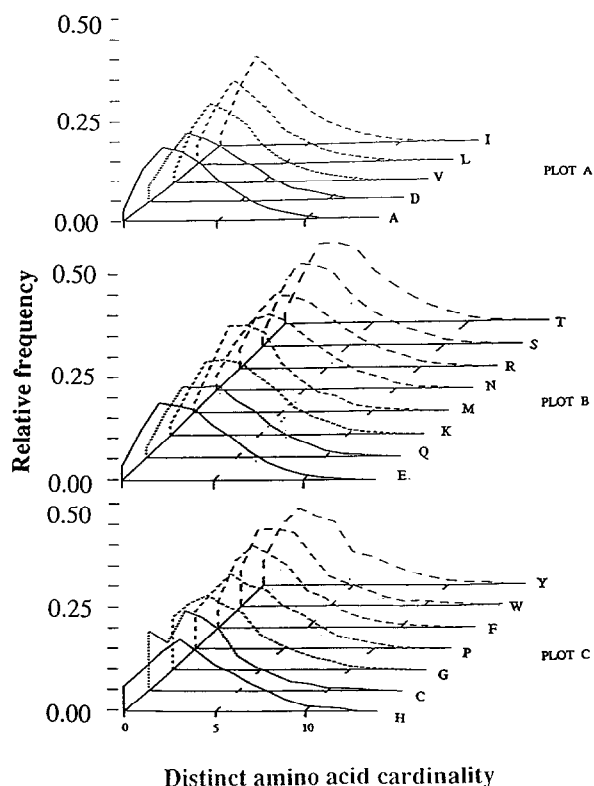


Fig. 2. A plot of the relative frequency of distinct amino acids in the aligned homologs for each amino acid in the set of 117 known structures. The position 0 on the horizontal axis indicates that the amino acid was common to all homologs, while 5 indicates that there were five distinct amino acids among the homologs, and so on. The amino acid homolog frequency distributions were grouped to emphasize similar distributions. The amino acids in plot A have a distribution differing little, if at all, from a Poisson. Those in plot C display the highest conservation (exhibit relatively large frequency at zero cardinality) and seem to be the furthest from a simple Poisson. Those in plot B are somewhat intermediate.

be in a β -strand than any other structure. On the other hand, 12% of the Gly residues are found conserved in aligned positions. While it is three times less likely to be found in a helix than anywhere else, it is ten times less likely to be found conserved in a helix and twenty-two times more likely to be found conserved in a loop or turn (including the ends of strands and helices)!

One way to estimate the potential utility of such homolog data is to define two classes of aligned positions, one containing single unique amino acids (conserved positions), and one containing two or more distinct amino acids. The corresponding prior probabilities are denoted by $P(a_k \& a_k')$ in Table I and their knowledge yields a small but clear gain in information. The extension of the amino acid alphabet to twenty-one to include the alignment gap character also gives some additional information (data not shown). The current data set is sufficient to calculate the conditional probabilities $P(S_k$

$|a_k \& a_k')$ and $P(S_k | S_{k-1} \& a_k \& a_k')$ as functions of all 380 possible homolog residue pairs. As the last two entries in Table I show, there is significant gain in information.

The conclusion of this study is not so much new, as it quantifies what has come to be fairly well accepted or recognized: that is, that local sequence information can provide estimates on the probable structure of a protein only up to a rather well-defined limit. However, this limit may be nearer to 85% than 65% if algorithms can be devised that exploit homolog information. Therefore the pursuit of secondary structure prediction for individual residues with accuracies in the high nineties requires additional information that is not available in any 'local sequence' statistical distillation of the known structures. This additional information could include the correlations between secondary structural elements imposed by the tertiary structure and might take the form of most probable secondary structure packing geometries and connectivities [8]. This in turn would require that we be able to describe those structures in such a manner that the relevant correlations (or constraints) can be recognized and quantified. An alternative approach is rejecting unlikely conformations instead of simply accepting the most likely one. This filtering approach does not generally yield a unique structure assignment for all residues, but the prediction success rate can be increased considerably [5].

Acknowledgements: The authors would like to thank Ilya Muchnik, Jean Garnier and Collin Stultz for their comments and suggestions, and Randall Smith for his help in organizing the data.

REFERENCES

- [1] Holley, L.H. and Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* 86, 152–156.
- [2] Qian, N. and Sejnowski, T.J. (1988) *J. Mol. Biol.* 202, 865–884.
- [3] Zhang, X., Mesirov, J.P. and Waltz, D.L. (1992) *J. Mol. Biol.* 225, 1049–1063.
- [4] Muskal, S.M. and Kim, S.H. (1992) *J. Mol. Biol.* 225, 713–727.
- [5] Vajda, S. (1993) *J. Mol. Biol.* 229, 125–145.
- [6] Greer, J. (1990) *Proteins* 7, 317–334.
- [7] Zvebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E. (1987) *J. Mol. Biol.* 195, 957–961.
- [8] Stultz, C.M., White, J.W. and Smith, T.F. (1993) *Protein Sci.* 2, 305–314.
- [9] Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987) in: *Crystallographic Databases: Information Content, Software Systems, Scientific Applications* (Allen, F.H., Bergerhoff, G. and Sievers, R. eds.) pp. 107–132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.
- [10] Smith, T.F. and Waterman, M.S. (1981) *Adv. Appl. Math.* 2, 482–489.
- [11] Smith, R.F. and Smith, T.F. (1992) *Protein Eng.* 5, 35–41.
- [12] Kabsch, W. and Sander, C. (1983) *Biopolymers* 22, 2577–2637.
- [13] Baierlein, R. (1971) *Atoms and Information Theory*, pp. 60–74, W.H. Freeman and Co., San Francisco.
- [14] Lim, V.I. (1974) *J. Mol. Biol.* 88, 873–894.